PageX: An Integrated Document Processing and Management Software for Digital Library

Hanchuan Peng, Zheru Chi, Wan-Chi Siu, and David D. Feng

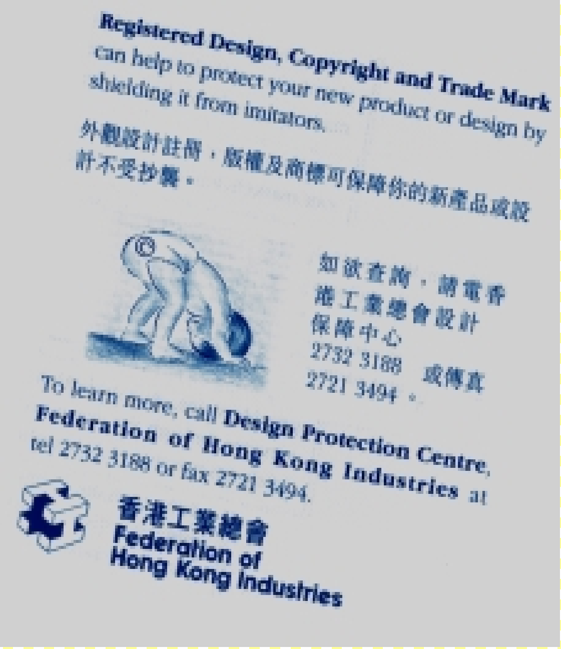
Department of Electronic & Information Eng.,
The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong.
Email: phc@eie.polyu.edu.hk

2000-Jan-10

Abstract

For digital libraries it is very important to design and to implement powerful engines to convert information on paper to electronic format. In this paper a PageX software is proposed as the integration of such engines, which include a set of intelligent document processing functions, a set of compact document management strategies, and a set of advanced accessories. With this software, a paper document will first be input as an optical image, which may be a mixture of graphics and text and may be skewed. The image will then be analyzed and decomposed into a series of component blocks, and encoded and stored in a structured and compact format. Well-developed accessory functions, including block editing and page annotation, page reconstruction and virtual editing, page matching and registration, document retrieval, etc., are provided to support advanced applications. With these carefully designed functions and strategies, PageX minimizes manual operations to a minimal degree.

Main Process of Working



Registered Design, Copyright and Trade Mark can help to protect your new product or design by shielding it from imitators.

外觀設計註冊，版權及商標可保障你的新產品或設計不受抄襲。

如欲查詢，請電香港工業總會設計保障中心 2732 3188 或傳真 2721 3494。

To learn more, call **Design Protection Centre, Federation of Hong Kong Industries** at tel 2732 3188 or fax 2721 3494.

香港工業總會
Federation of Hong Kong Industries

Image Acquiring Devices

PageX

Binarization
Skew correction
Text blocks extraction
Script determination
Character segmentation
Thinning
Feature extraction
……

Page Format Database

PageX

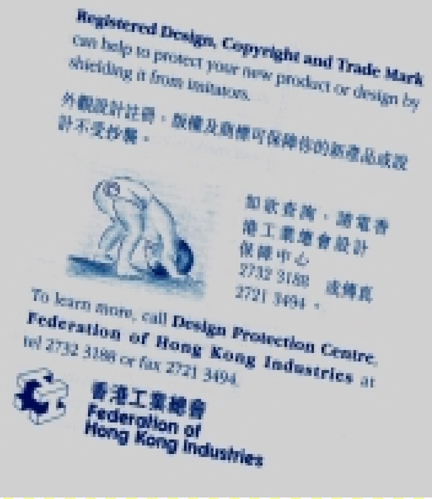
Form Registration

PageX

Character Recognition
Page/Form Reformatting
Unrecognizable Block/Image Coding
Page Synthesis/Reconstruction
……

Main Structure of the System

Optical Images



Registered Design, Copyright and Trade Mark can help to protect your new product or design by shielding it from imitators.

外觀設計註冊，版權及商標可保障你的新產品或設計不受抄襲。

如欲查詢，請電香港工業總會設計保障中心2732 3188或傳真2721 3494。

To learn more, call **Design Protection Centre, Federation of Hong Kong Industries** at tel 2732 3188 or fax 2721 3494.

香港工業總會
Federation of Hong Kong Industries

Engine Set I: Page Analysis and Decomposition

(1) Foreground Extraction and Binarization
(2) Correction of Unknown Skew Angle
(3) Page Blocking
(4) Language Separation
(5) Character Separation
(6) Character Recognition
(7) ……

Component Block List

Engine Set II: Block Coding and Page Management

(1) Block Compression and Decoding
(2) Page Layout Control
(3) Dynamically Adjustable Data Structures
(4) ……

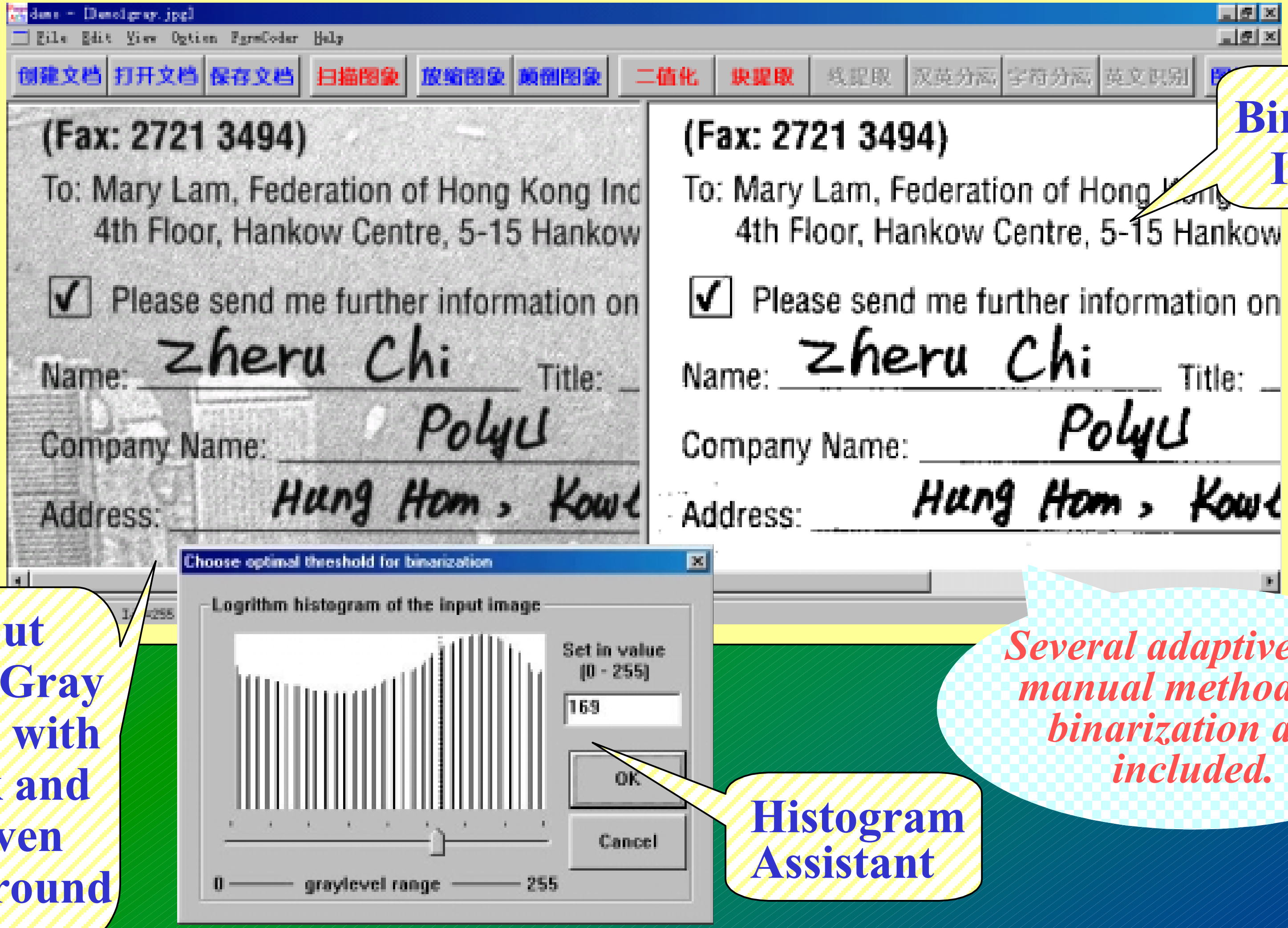
Compact e-Doc

Engine Set III: Advanced Accessories

(1) Virtual Editing
(2) Page Matching
(3) Database Linking
(4) Retrieval
(5) ……

Compact e-Doc

Binarization


demo - [Demolgray.jpg]
File Edit View Option FgrmCoder Help
创建文档 打开文档 保存文档 扫描图象 放缩图象 颠倒图象 二值化 块提取 线提取 汉英分离 字符分离 英文识别

(Fax: 2721 3494)

To: Mary Lam, Federation of Hong Kong Ind
4th Floor, Hankow Centre, 5-15 Hankow

☑ Please send me further information on

Name: Zheru Chi Title:

Company Name: PolyU

Address: Hung Hom, Kowl

(Fax: 2721 3494)

To: Mary Lam, Federation of Hong Kong
4th Floor, Hankow Centre, 5-15 Hankow

☑ Please send me further information on

Name: Zheru Chi Title:

Company Name: PolyU

Address: Hung Hom, Kowl

Choose optimal threshold for binarization

Logrithm histogram of the input image

Set in value [0 - 255]: 169

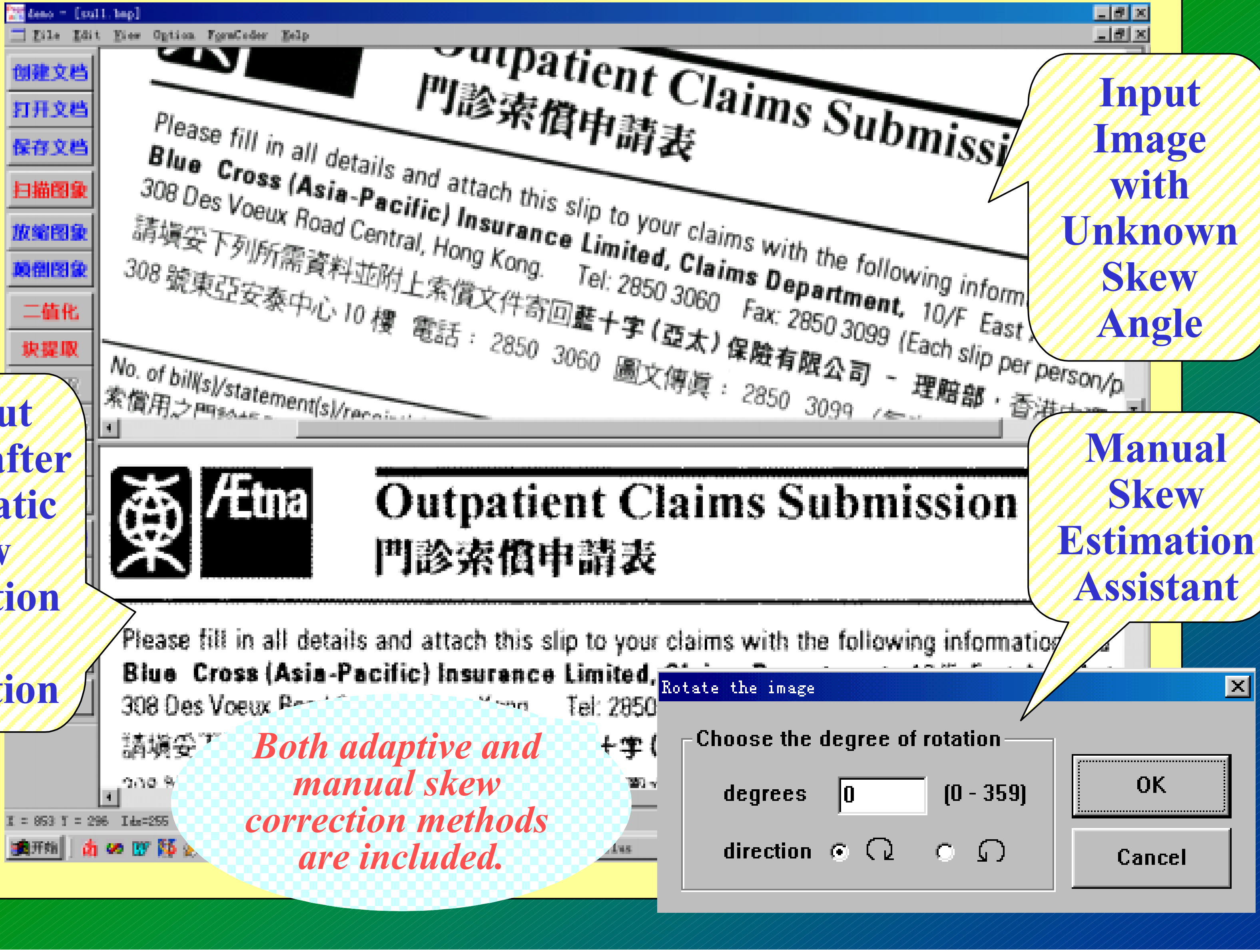
Binarized Image

Input Color/Gray Image with Thick and Uneven Background

Histogram Assistant

Several adaptive and manual methods of binarization are included.

Skew Correction



Outpatient Claims Submissi

門診索償申請表

Please fill in all details and attach this slip to your claims with the following inform

Blue Cross (Asia-Pacific) Insurance Limited, Claims Department, 10/F East

308 Des Voeux Road Central, Hong Kong. Tel: 2850 3060 Fax: 2850 3099 (Each slip per person/p

請填妥下列所需資料並附上索償文件寄回藍十字(亞太)保險有限公司 - 理賠部，香港中環

308 號東亞安泰中心 10 樓 電話： 2850 3060 圖文傳真： 2850 3099 （每

No. of bill(s)/statement(s)/receip

索償用之門診

Outpatient Claims Submission

門診索償申請表

Please fill in all details and attach this slip to your claims with the following informatio

Blue Cross (Asia-Pacific) Insurance Limited,

308 Des Voeux R

Tel: 2850

請填妥

Rotate the image

Choose the degree of rotation

degrees 0 [0 - 359]

direction ● ○

OK

Cancel

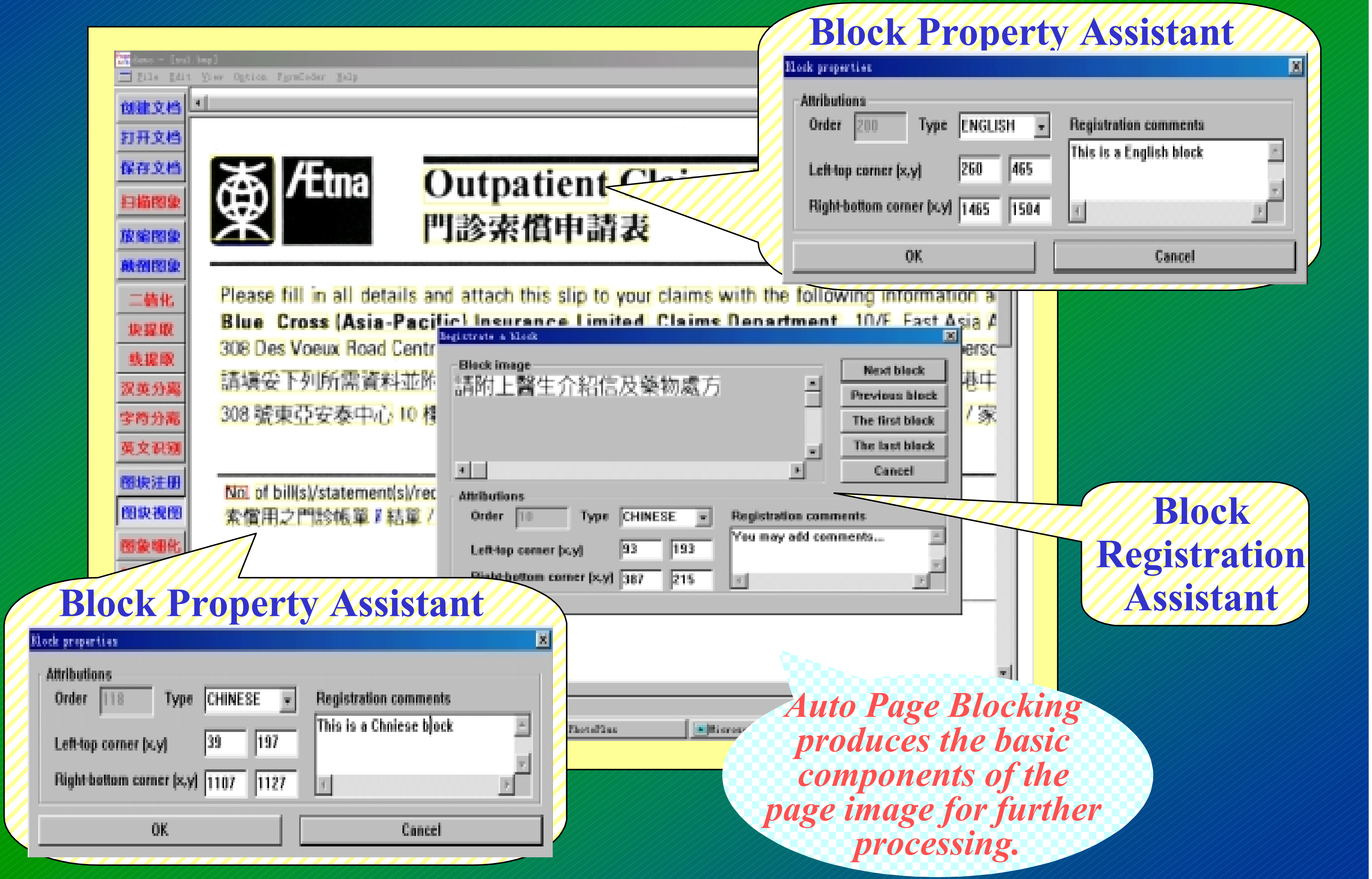
Input Image with Unknown Skew Angle

Output Image after automatic skew estimation and correction

Manual Skew Estimation Assistant

Both adaptive and manual skew correction methods are included.

Page Blocking



Outpatient Clai

門診索償申請表

Please fill in all details and attach this slip to your claims with the following information a

Blue Cross (Asia-Pacific) Insurance Limited, Claims Department, 10/F, East Asia A

308 Des Voeux Road Centr

請塡妥下列所需資料並附

308 號東亞安泰中心 10 樓

No. of bill(s)/statement(s)/rec

索償用之門診帳單 / 結單 /

Block Property Assistant

Block properties

Attributions

Order: 200 Type: ENGLISH

Left-top corner (x,y): 260 465

Right-bottom corner (x,y): 1465 1504

Registration comments: This is a English block

Registrate a block

Block image: 請附上醫生介紹信及藥物處方

Attributions

Order: 18 Type: CHINESE

Left-top corner (x,y): 93 193

Right-bottom corner (x,y): 387 215

Registration comments: You may add comments...

Block Registration Assistant

Block Property Assistant

Block properties

Attributions

Order: 118 Type: CHINESE

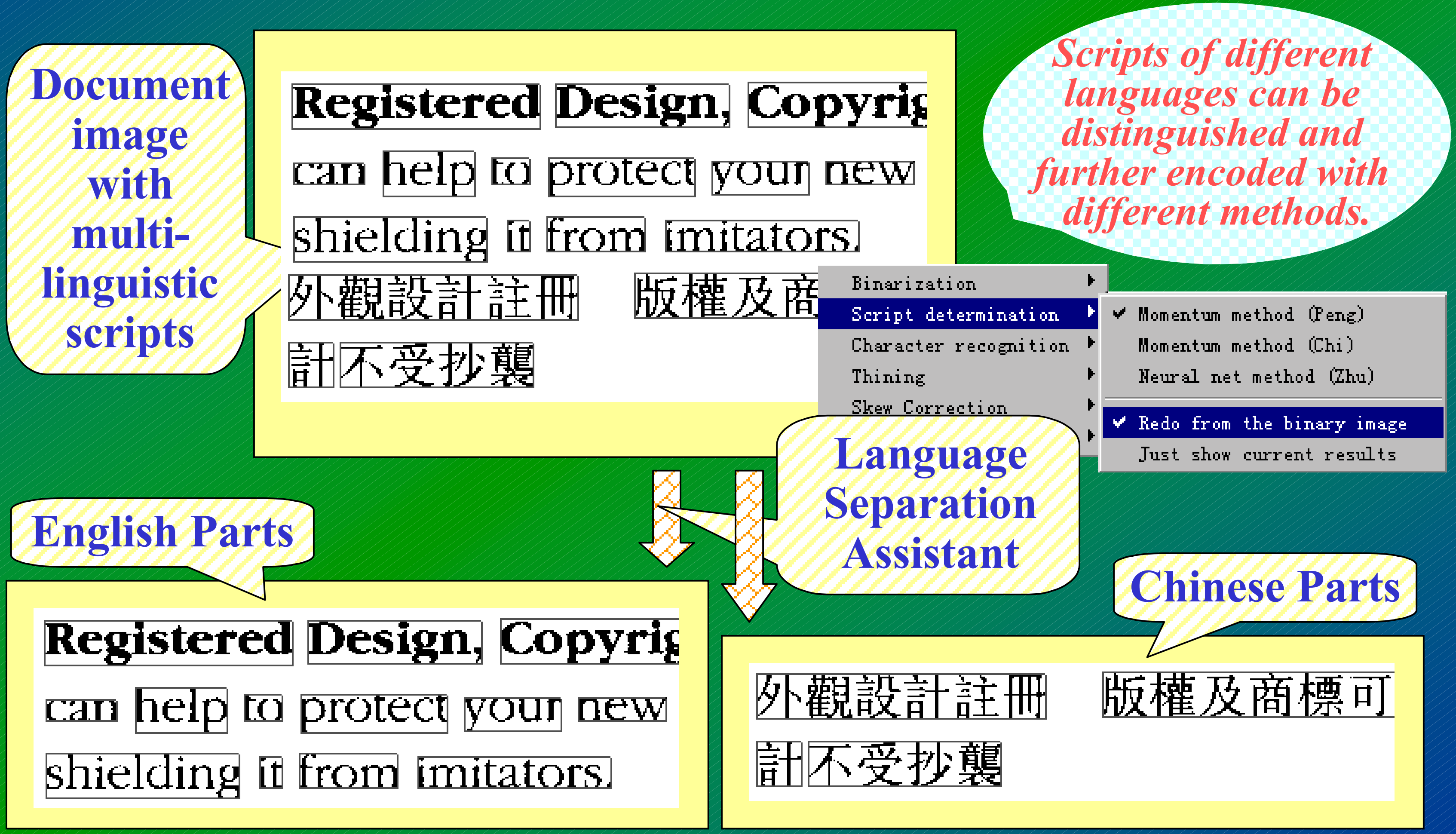
Left-top corner (x,y): 39 197

Right-bottom corner (x,y): 1107 1127

Registration comments: This is a Chniese block

Auto Page Blocking produces the basic components of the page image for further processing.

Script Determination



Document image with multi-linguistic scripts

Registered Design, Copyrig
can help to protect your new
shielding it from imitators.
外觀設計註冊 版權及商
計不受抄襲

Scripts of different languages can be distinguished and further encoded with different methods.

Language Separation Assistant

English Parts

Registered Design, Copyrig
can help to protect your new
shielding it from imitators.

Chinese Parts

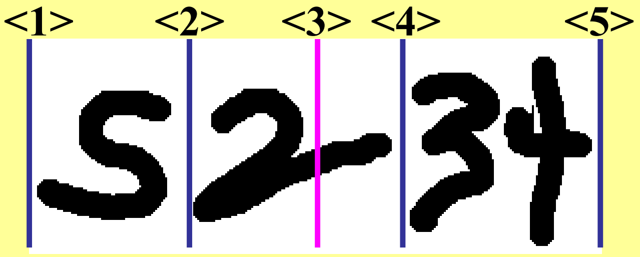
外觀設計註冊 版權及商標可
計不受抄襲

2000-Jan-10

Virtual Editing

The character image and text are manipulated in a unified way.

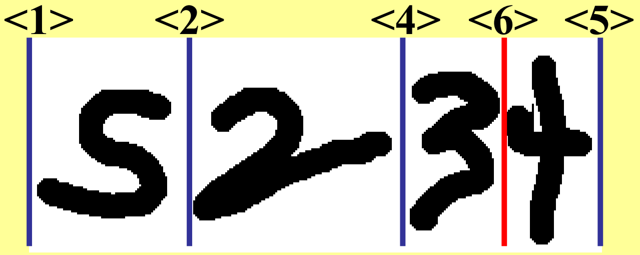
Wrongly separated text image



S7-汁

Wrongly recognized text

Text image after virtual editing

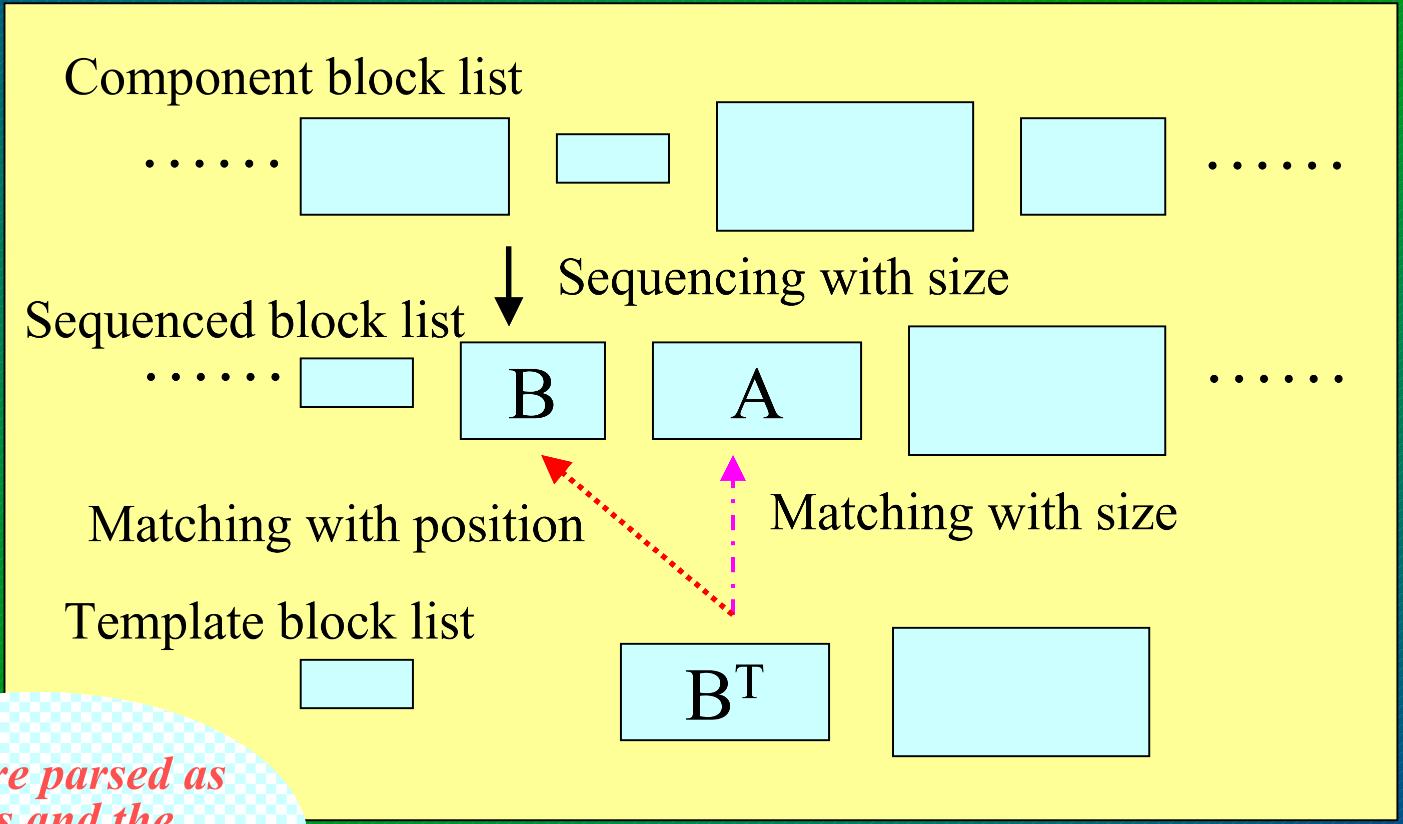


5234

Correctly recognized text

2000-Jan-10

Page Matching



Pages are parsed as blocks and the block sequences are matched.

2000-Jan-10